基于流形学习的新闻主题关系构建和演化研究*

徐月梅[1]　李　杨[2,3]　梁　野[1]　蔡连侨[1]
[1](北京外国语大学计算机系　北京　100089)
[2](中国科学院信息工程研究所　北京　100093)
[3](中国科学院大学　北京　100049)

摘要: 【**目的**】通过对以互联网为媒介的新闻报道的主题演化研究, 分析新闻主题的产生、发展和演变过程, 把握媒体舆论方向。【**方法**】引入流形学习构建全局时间跨度的新闻主题关联关系, 挖掘由 LDA 主题模型识别得到的各个时间窗口的高维主题向量间的关系, 在低维平面上实现主题聚类和相互关联的可视化, 提出利用社会网络理论指标分析主题的演化结果。【**结果**】利用 2015 年美国有线电视新闻网对中国的新闻报道进行主题关系构建和演化, 结果表明该方法能够发现主题在全局时间跨度的演化趋势。【**局限**】时间窗口长度对主题演化的效果和可变时间窗口长度机制没有涉及。【**结论**】新闻主题演化分析方法能够在低维可视平面上描绘主题在全局时间跨度的演化, 避免主题由于相邻时间窗口关联失效而导致全局演化路径的断裂。

关键词: 潜在狄利克雷分配模型　流形学习　主题关联　主题演化
分类号: TP393　G354

1　引　言

随着信息技术的发展, 互联网已成为信息传播的重要渠道, 被公认为是继报纸、广播、电视之后的"第四媒体"[1]。研究以互联网为媒介的西方主流媒体对中国的新闻报道, 有助于了解西方媒体中的中国形象, 把握国外舆论的发展方向。新闻报道的主题演化是指新闻报道的主题内容与强度在研究过程中变化的现象, 一般经历从提出、发展、衰亡到最后结束的过程。例如天津塘沽大爆炸事件, 美国主流媒体有线电视新闻网(Cable News Network, CNN) 2015 年 8 月 13 号第一次进行报导, 随后在 14-21 号每天都有相关新闻追踪, 而 27 号是最后一次报导, 意味着该事件主题的结束。可见, 随着时间的变化, 西方媒体对中国的新闻报道主题也随着变迁, 如何描述新闻主题的演变过程是目前研究的难点[2]。

潜在狄利克雷分配(Latent Dirichlet Allocation, LDA)模型[3]是模拟文档生成过程的主题模型, 其参数空间的规模与文档数量无关, 适合处理大规模语料, 因此近年来成为主题演化研究的重要途径之一。常见的思路是利用 LDA 模型获取不同时间段的主题及其关键词, 将相邻时间窗口的主题根据关键词的近似程度采取阈值法[4]或最大相似度法[5]进行关联, 再从相邻时间窗口建立的主题关联关系观察多个时间窗口的主题演变。

然而, 基于相邻时间窗口的主题演化分析方法不能直接应用于新闻报道的主题演化, 原因有两点。首先, 基于相邻时间窗口的主题演变关系建立容易因为

通讯作者: 徐月梅, ORCID: 0000-0002-0223-7146, E-mail: xuyuemei@bfsu.edu.cn。

*本文系国家社会科学基金重大委托项目"语言大数据挖掘与文化价值发现"(项目编号:14@ZH036)、北京市社会科学基金研究基地项目"北京对外文化传播过程中'两微一端'影响力比较研究"(项目编号:15JDZHC011)和中央高校基本科研业务费专项资金资助项目"对外传播过程中互联网用户行为特征和影响力研究"(项目编号:023600-500110002)的研究成果之一。

某个相邻窗口的主题关联出错而导致整个主题链的演变失效。例如，某个主题的演变经过时间窗口[t1, t2, t3, t4]，但由于在[t2, t3]相邻时间窗口内该主题的关联出错(可能由于阈值设置过大或者相似度计算有误)使得该主题的全局演变过程断裂。其次，新闻报道的主题具有突发性和时间间隔性，使得新闻主题的演化规律并不一定遵循相邻时间窗口跨度。例如 2015 年 6 月 CNN 网站针对中国南海问题进行相关报道，在间隔 7 月、8 月之后，9 月和 10 月又有中国南海问题的相关报道，可见新闻报道主题的演化时间跨度具有不确定性。

针对上述两个问题，本文提出将流形学习(Manifold Learning)[6]引入到新闻主题的关系构建和演化研究，定义新闻主题的演化关系并不局限于传统的相邻时间间隔的主题演化，而是从全局时间跨度分析两个主题的关联关系。通过从整体上对各个时间窗口内的主题进行关联分析，以期获得主题在全局时间上的演变关系。经过 LDA 模型抽取得到的主题表现为高维度的特征词向量，采用现有的相似度计算方法进行全局时间上的主题关联因为"高维数灾难"[7]而变得十分困难。例如有 5 个时间窗口，每个时间窗口有 10 个主题，每个主题的向量维度为 1 000 维，利用相似度方法进行全局时间上的主题关联需要 4×10^7 时间复杂度($(5\text{-}1)\times10\times10^3\times10^3$)。而流形学习技术可以挖掘高维主题向量之间隐藏的关联关系，找到高维空间中的低维流形，并求出主题在相应的低维空间的嵌入映射，实现维数约简和可视化，使得进一步利用社会网络分析相关指标分析主题演变规律变为可能。本文的创新点与贡献总结如下：

(1) 借鉴图像处理和机器学习领域中的非线性降维思想，引入流形学习方法挖掘由 LDA 模型抽取得到的各个时间窗口的高维主题向量，一方面在低维平面上可视化高维主题向量间的关系，另一方面将非线性降维的结果与余弦相似度结合，重新定义低维平面上主题之间的距离，实现全局时间窗口的主题关联。

(2) 高维主题向量经非线性降维后表现为一个小型的社会网络：主题表征为低维平面上的节点，节点的远近反映主题之间的距离，节点的边为主题的关联边。因此利用社会网络理论的 4 种度量指标来分析主题的演化，识别主题演化过程中影响力大的主题、活跃的主题以及主题演化网络的整体属性等。

(3) 以美国 CNN 网站对中国的相关新闻报道为例验证了所提方法的有效性和准确性。

2 相关研究工作

早期对主题演化的研究主要是将文档的时间信息引入到 LDA 模型或其变形模型中，并利用连续的时间信息指导文档集中主题的分布，如连续时间模型 TOT[8]、动态主题模型 DTM[9]。但该方法无法对新文档进行扩展，新文档加入后必须重新建模。

近年来对主题演化研究主要有两种思路：一种是先对整个文档集合运用 LDA 获取主题，再从时间上将主题划分为各个子集，分析主题在各个子集上的分布从而得到主题的演化规律[10]。另一种是先对整个文档集合按照时间信息离散到各个时间窗口，再利用 LDA 获取各个时间窗口内的主题，最后将相邻时间窗口的主题关联，得到主题演化过程[2,4-5,11]。这两种方法各有其局限性。前一种方法依赖于时间粒度的选取，时间粒度的取值直接影响演化的准确性。后一种方法中，相邻时间窗口的主题关联是分析主题演化的重要步骤，不同的关联方法将得到不同的演化结果。例如，楚克明等[2]通过计算相邻时间段中任意两个主题的特征向量相似度实现主题关联度分析，该方法对阈值大小比较敏感并且阈值的确定需要较强的专业知识。崔凯等[11]使用 Kullback Leibler 相对熵来计算主题的相似性从而建立关联，但得到的主题演化都是一对一的，与科学研究中主题的融合、交叉等现象不完全吻合。

此外，为了提高阈值法或相似度法的主题关联准确性，相关文献提出了特征词过滤[12]和主题关联过滤的方法[13]。由阈值法或相似度法建立主题关联后，定义过滤规则去除无效的关联来提高主题关联的准确性，但其效果的提高过度依赖于过滤规则的定义，过滤规则对于不同领域的主题不具有普适性。

总体而言，不管是先获取主题再从时间上划分子集分析主题演化，还是先划分时间窗口再获取主题从而得到主题演化，现有研究都是从相邻时间窗口构建主题的演化关系。一方面容易因为相邻窗口内的主题关联出错使得全局演化过程断裂；另一方面新闻报道主题演化的时间跨度具有随机性，不一定遵循相邻时间窗口的跨度。

为了解决上述两个问题，本文从一个全新的角度，

首次引入流形学习方法从全局时间跨度、而非相邻时间窗口跨度构建新闻的主题关系，并利用社会网络相关分析指标衡量主题演化的结果。流形学习近年来被广泛应用在数据挖掘、机器学习、模式识别等领域，其作为解决非线性降维问题的方法，在挖掘高维数据集的固有特征分布和结构特点方面具有优势[6]。经过LDA 抽取的主题表征为非线性、高维度的特征词向量，若采用现有的相似度计算方法将由于“高维数灾难”难以进行全局时间跨度的主题关联，而流形学习能够挖掘高维度主题向量之间蕴含的关联关系，将其映射到低维空间，使得全局时间跨度的新闻主题关系构建和主题演化分析变为可能。

3　基于流形学习的新闻主题关系构建和演化分析

3.1　基本思路

本文提出的基于流形学习的新闻主题演化关系构建和演化方法的基本流程如图 1 所示:

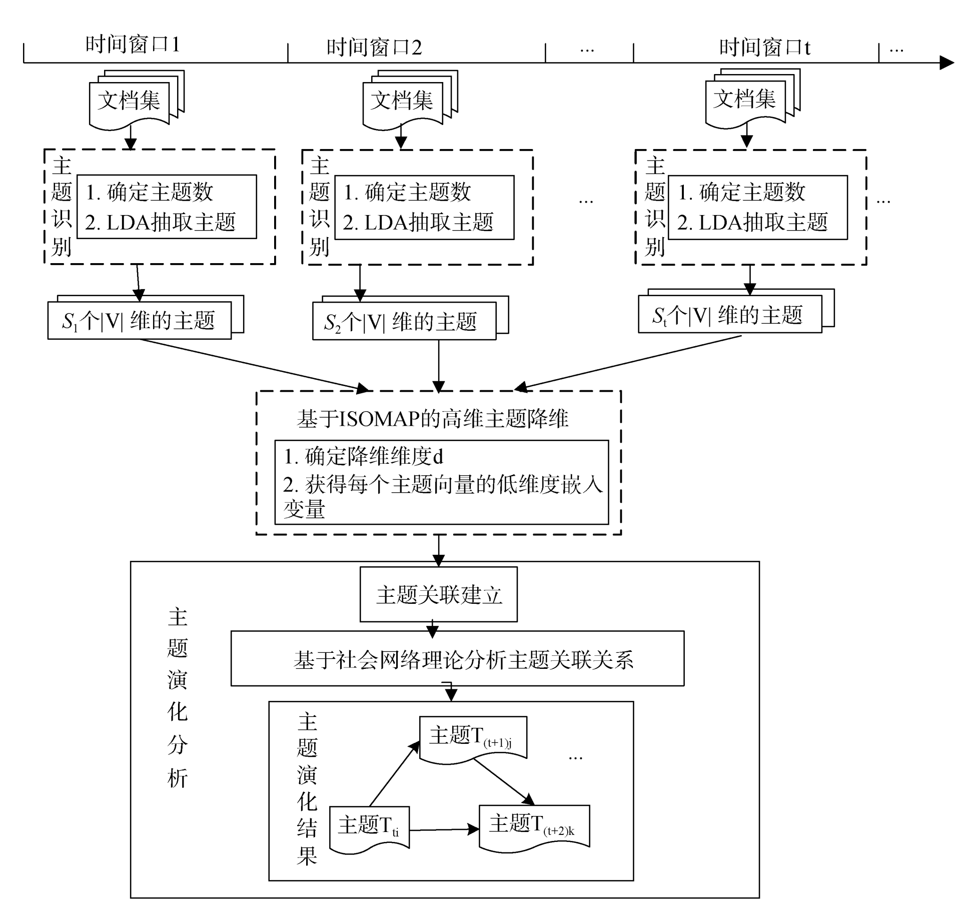


图 1　基于流形学习的新闻主题演化方法流程图

(1) 将时间序列划分为若干个长度固定的时间窗口，根据时间将文本划入到相应的时间窗口，利用LDA 模型抽取每个时间窗口的主题，并将主题表示为高维特征词向量的形式。

(2) 将得到的多个高维主题向量利用流形学习算法进行非线性降维，获得每个主题向量的低维度嵌入变量以及主题关联边。

(3) 为主题关联边赋予权重，确定主题间的关联关系。

(4) 利用社会网络指标分析主题关联关系，分析主题的演化特征。

(5) 得到新闻主题演化结果。

3.2　新闻主题的定义和识别

新闻报道的主题表现为媒体对某一特定事件及其

所有相关事件的集合(简称主题)。给定 D 个新闻报道文本，用集合 $C=\{d_1,d_2,\cdots,d_D\}$ 表示。V 为所有文本不相同单词构成的词汇集合。将主题定义为一组语义上相关的词及词语主题相关的权重的向量表示[13]:

$$T=\{(v_1,p_1),(v_2,p_2),\cdots,(v_k,p_k),\cdots,(v_{|V|},p_{|V|})\} \quad (1)$$

其中，$v_k \in V$ 是与主题 T 相关的词，p_k 是主题 T 在该词上的分布概率。

将时间序列划分为 n 个长度为 L 的时间窗口，依据时间将集合 C 中新闻报道划分到相应的时间窗口，C_t 表示时间窗口 t 的新闻报道集合。采用 LDA 模型对 $C_t, t\in[1,n]$ 抽取主题。LDA 模型是一个三层贝叶斯文本主题生成模型，可以发现任何离散数据中潜在的主题结构。其基本思想是：假设文档由若干个潜在主题的混合组成，而每个主题由若干个词的分布刻画。

LDA 设立参数α作为文本集合的主题先验超参数，β为主题集合的词汇先验超参数，使得每篇文本服从参数为α的 Dirichlet 分布，每个主题服从参数为β的 Dirichlet 分布。给定文本集合，根据 Gibbs 采样[14]计算出文本–主题概率分布 θ 和主题–词分布 φ 如下:

$$\theta_{mi}=\frac{n_m^{(i)}+\alpha}{\sum_{j=1}^{|S|} n_m^{(j)}+|S|\alpha} \quad m\in[1,D], i\in[1,|S|] \quad (2)$$

$$\varphi_{ik}=\frac{n_i^{(k)}+\alpha}{\sum_{j=1}^{|V|} n_i^{(j)}+|V|\alpha} \quad i\in[1,|S|], k\in[1,|V|] \quad (3)$$

其中，θ_{mi} 为文本 d_m 属于主题 T_i 的概率，$n_m^{(i)}$ 表示文本 d_m 中赋予主题 T_i 的词的总数。φ_{ik} 为主题 T_i 出现单词 v_k 的概率，$n_i^{(k)}$ 表示词 v_k 被赋予主题 T_i 的总次数。S 为 LDA 抽取的主题集合。

结合公式(1)的定义和 LDA 模型，笔者将时间窗口 t 内文档集合 C_t 的主题表示为:

$$T_{ti}=\{(v_1,\varphi_{ti1}),(v_2,\varphi_{ti2}),\cdots,(v_k,\varphi_{tik}),\cdots,(v_{|V|},\varphi_{ti|V|})\} \quad (4)$$

其中，$1\leqslant i\leqslant S_t$，$S_t$ 为时间窗口 t 内的主题数目，T_{ti} 的向量维度为 $|V|$ 维。$v_k\in V$，φ_{tik} 由 LDA 模型计算得到，为主题 T_{ti} 出现单词 v_k 的概率。

每个时间窗口内的新闻报道数不同，相应的主题数也随之动态变化。S_t 的最佳值采用统计语言模型中常用的评价标准——困惑度(Perplexity)[15]进行选取，计算如下:

$$Perplexity(C_t)=\exp\{-\frac{\sum_{m=1}^{|C_t|}\ln P(d_m)}{\sum_{m=1}^{|C_t|}N_m}\} \quad (5)$$

其中，N_m 表示第 m 篇新闻报道的长度，$P(d_m)$ 表示模型产生第 m 篇新闻报道的概率。困惑度的值越小，性能越好。在其他参数确定的情况下，通过对 S_t 取不同值进行困惑度的计算和分析，选取得到最优主题数目的 S_t 值。

对 n 个时间窗口分别抽取主题，笔者将总的主题集合 TopicSet 以及总主题数 S 定义为:

$$TopicSet=(T_{11},T_{12}\cdots;\cdots;T_{t1,}T_{t2}\cdots;T_{n1},T_{n2}\cdots) \quad (6)$$

$$S=\sum_{t=1}^{n}S_t \quad (7)$$

3.3 基于流形学习的主题演化关系构建

主题演化反映了主题变化的过程，主题的演化在时间上存在延续性。对 n 个时间窗口的文本经 LDA 识别，得到的主题表现为 S 个 $|V|$ 维的特征词向量，当 $|V|$ 较大时使得挖掘主题间的演化关系变得困难。本文利用流形学习对高维度的主题向量进行降维，并构建主题演化关系。

流形学习是一种非线性降维方法，可用于处理高维数据，通过对高维空间的特征数据学习以获得低维的隐变量模型，即找到高维空间中的低维流形，以实现维数约简和可视化。图 2 展示了高维流形与低维映射的关系，在三维空间中的"瑞士蛋卷"数据分布模型，经过降维后在二维平面上显示各个数据点的关系[16]。

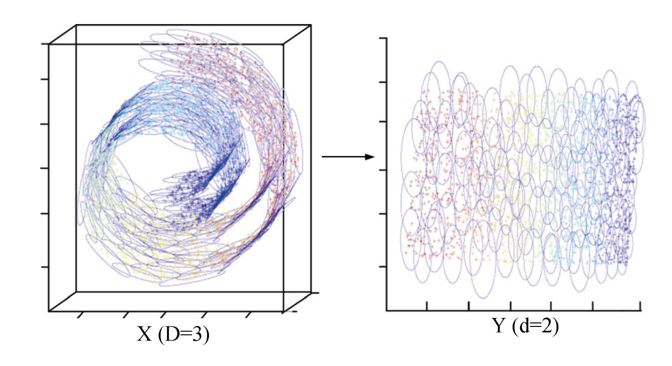


图 2　高维流形向低维空间的映射

流形学习的典型实现方法包括等距特征映射(Isometric Feature Mapping, ISOMAP)[17]和局部线性嵌入(Locally Linear Embedding, LLE)[18]等。本文采用

ISOMAP 算法，该算法主要思想是利用局部邻域的欧氏距离近似计算数据点之间的全局流形测地线距离，通过建立原数据之间的测地线距离与降维数据间的空间距离的对等关系从而实现降维。ISOMAP 在降维过程中通过计算点与点之间的测地距离，并采用多维标度法(Multi-Dimensional Scaling, MDS)[17]来获取全局最优的几何结构，从而准确发现数据流形潜在的参数空间。

为了得到高维主题向量 T_{ti} 的特征，需要在欧式空间 R^d 找到一个低维度区域 Y_{ti} 反映 $T_{ti} \in R^{|V|}$ 的特征，通常 $d << |V|$。笔者将主题 T_{ti} 在欧式空间 R^d 的低维度嵌入变量 Y_{ti} 定义为：

$$Y_{ti} = \{y_{ti}(1), y_{ti}(2), \cdots, y_{ti}(d)\} \mid Y_{ti} \in R^d \quad (8)$$

其中，d 是 Y_{ti} 的维度。算法 1 将上述 S 个 $|V|$ 维 ($|V| >> 1$) 的主题向量集合 TopicSet 进行基于 ISOMAP 的高维主题向量降维。取维度空间 d 为 2，笔者将 TopicSet 在二维平面的嵌入变量定义为：

$$(Y_{11}, Y_{12} \cdots; \cdots; Y_{t1}, Y_{t2} \cdots; Y_{n1}, Y_{n2} \cdots) \quad (9)$$

其中，Y_{ti} 为 T_{ti} 的低维嵌入变量，可在二维平面显示，有利于直接观察主题之间的演化关系。

算法 1　基于 ISOMAP 的高维主题降维算法

输入：TopicSet；

输出：每个主题向量 T_{ti} 的低维度嵌入变量 Y_{ti} 和 Y_{ti} 的邻域图邻接矩阵 E；

执行：

①建立每个主题 T_{ti} 的邻域图。

根据主题向量之间的距离，确定主题集合中哪些主题为邻居主题。计算所有主题之间的欧氏距离 $d_T(i,j)$，确定每个主题的 K 个最近主题，K 为可输入变量。这些主题的邻居关系被描述在一个覆盖采样点的带权图 G 中，主题之间的关系以链路权重 $d_T(i,j)$ 表示。

②计算图 G 中主题之间的测地线距离。

根据步骤①确定的图 G 和两两主题之间的链路权重 $d_T(i,j)$，计算所有主题之间的最短路径 $d_G(i,j)$，并以此来估算流形内所有主题之间的测地线距离。

③构建低维度的嵌入变量 Y_{ti} 和 Y_{ti} 的邻域图邻接矩阵 E。

对于步骤②得到的所有主题之间的最短路径距离矩阵 $D_G = \{d_G(i,j)\}$，应用多维标度法进行降维，创建位于 d-维欧氏空间内的低维嵌入变量 Y_{ti} 和 Y_{ti} 的邻域图邻接矩阵 E。

降维后得到每个主题的低维度嵌入变量 Y_{ti} 和低维度嵌入变量的邻接矩阵 E。其中，$Y_{ti} = (x_{ti}, y_{ti})$，$x_{ti}$ 和 y_{ti} 为主题 Y_{ti} 在二维平面的横坐标和纵坐标值，E 为 0-1 矩阵。每个主题 T_{ti} 表征为二维平面上的一个点，节点在二维平面上的分布由高维主题向量的测地线距离决定，反映了主题之间的相似程度。节点越密集表示具有演化关系的相似主题越多，反之则越少。为了建立全局时间跨度的主题关联，笔者基于余弦相似度[19]，在二维平面上重新定义任意两个时间窗口内的主题距离为：

$$Sim(Y_{ti}, Y_{(t+k)j}) = \begin{cases} \dfrac{x_{ti}x_{(t+k)j} + y_{ti}y_{(t+k)j}}{\sqrt{x_{ti}^2 + y_{ti}^2} \times \sqrt{x_{(t+k)j}^2 + y_{(t+k)j}^2}} & E(I_{ti}, I_{(t+k)j}) = 1 \\ 0 & E(I_{ti}, I_{(t+k)j}) = 0 \end{cases} \quad (10)$$

其中，Y_{ti} 和 $Y_{(t+k)j}$ 分别为时间窗口 t 和 t+k 内的主题低维度嵌入变量，$i \in S_t, j \in S_{t+k}$，$t \in [1, n-1]$，$k \geqslant 1$。$E(I_{ti}, I_{(t+k)j}) = 1$ 表示主题向量 Y_{ti} 和 $Y_{(t+k)j}$ 在低维嵌入平面上有关联边，反之则表示主题向量 Y_{ti} 和 $Y_{(t+k)j}$ 关联程度低，将其相似度赋值为 0。

3.4　社会网络指标的主题演化分析

一个社会网络由多个点和各点之间的连线组成，“点”是各个社会行动者，“边”是行动者之间的各种社会关系。高维的主题特征向量经过 ISOMAP 降维表现为一个小型的社会网络：由主题节点之间的相互作用关系形成的二维平面图。其中，二维平面上节点之间的距离表征主题之间的关系和相互作用程度。因此，可借鉴社会网络理论的 4 种度量指标[20]来分析主题的演化，识别演化过程中影响力大的主题、活跃主题和主题演化网络的整体属性等：

(1) 度(Degree)，以连接到节点的边的数目作为度量节点重要性的依据。在有向图中，节点的度包括点入度和点出度。在主题构成的有向图中，如果一个主题拥有更高的度数值，则该主题与很多其他主题存在演变关系。其中，入度值越高，则在演变过程中有越多主题指向到该主题；出度值越高，则该主题有越多延续主题。度数仅仅描述主题所产生的局部影响力，无法反映主题的全局演变情况。

(2) 介数中心度(Betweeness Centrality)，以网络中经过该节点的所有点与点的最短路径的数目作为度量依据。介数中心度反映节点的信息交互能力，可用来衡量一个主题作为媒介者的能力，即占据在其他两个主题演变路径之间的交互能力。在主题的演变分析中，

通过介数中心度，可以确定比较活跃的主题。

(3) 密度(Density)，是一个网络图中实际存在的边数与可能存在的最多边数的比值，一般用来衡量网络图的全局凝聚力水平。在主题构成的网络图中，密度越大则主题的演变关系越复杂，演化关系越多；密度越小则主题的演变关系越简单，演化关系越少。

(4) 直径(Diameter)，将网络中最长测地线的长度作为度量依据，测地线是给定两点之间最短的路径。在主题演变图中，存在多条测地线，而直径表征主题演变关系上最长的演变距离跳数。

4 实 验

为了验证基于流形学习的新闻主题关系构建和演化分析方法的有效性，实验基于 GooSeeker 数据爬取平台[21]从 CNN 网站抓取了 2015 年与中国相关的新闻报道作为文本集，共 464 篇新闻报道。对文本集的每一篇文档进行数据预处理，包括分词、剔除停用词、词形还原、词干提取、提取关键词等。

将时间序列划分为 12 个长度为 1 个月的时间窗口，根据新闻的报导时间将其划入到相应的窗口。表 1 列举了各时间窗口的新闻报道数以及利用公式(5)确定各个时间窗口的最优主题数。

表 1 数据集各时间窗口所含新闻报道数和最优主题数

新闻报道集	文档数	最优主题数
2015 年 1 月	27	5
2015 年 2 月	16	5
2015 年 3 月	21	4
2015 年 4 月	25	6
2015 年 5 月	41	5
2015 年 6 月	38	6
2015 年 7 月	71	7
2015 年 8 月	72	6
2015 年 9 月	66	6
2015 年 10 月	24	6
2015 年 11 月	34	5
2015 年 12 月	29	5
总计	464	66

4.1 主题识别结果

利用LDA模型抽取每个时间窗口的主题，设置两个超参数为 $\alpha = 50/L$，$\beta = 0.01$ [3]。选取每个主题中分布概率 Top20 的单词作为主题内容的特征词。表 2 列举了抽取的部分主题(并给出了人工总结后的主题内容)及其特征词(仅列举前 10 个)。

表 2 2015 年 CNN 与中国相关的部分主题

主题	主题内容	主题特征词(前 10 个)
T_{53}	南海 军事	sea, south, island, military, navy, aircraft, flight, state, surveillance, warn
T_{64}	南海 袭击	government, attack, state, sea, island, official, hack, federal, information, south
T_{104}	南海 领土	island, sea, operation, reef, south, water, freedom, beijing, dispute, territorial
T_{114}	习近平与马英九会面	taiwan, ma, xi, meeting, beijing, president, relation, state, Singapore, mainland
T_{23}	希腊 经济	Greece, bank, currency, russia, internet, growth, financial, government, economist, money
T_{31}	柴静 空气污染	state, video, chai, government, xi, president, air, pollution, authority, documentary
T_{42}	市场 股票	state, investor, government, market, stock, growth, global, charge, unite, economic
T_{71}	股票 崩盘	market, stock, economy, share, shanghai, financial, investor, trade, government, crash

从表 2 可看出：LDA 模型能够识别每个时间窗口内的新闻报道主题，主题的类别包括军事(T_{53}、T_{64}、T_{104})、政治(T_{114})、经济(T_{23}、T_{42}、T_{71})和社会民生(T_{31})等方面。各主题中分布概率较高的主题特征词能够涵盖该主题的内容。以 5 月份的第 3 个主题为例(T_{53})，该主题与中国南海军事主权有关，Top10 的特征词为：sea(海洋)，south(南方)，island(岛屿)，military(军事)，navy(海军)，aircraft(航空器)，flight(飞行)，state(声明)，surveillance(监督)，warn(警告)。

4.2 基于 ISOMAP 流形学习的主题关联结果

根据 3.3 节中叙述的方法，进行基于 ISOMAP 流形学习的主题关联分析。每个主题选取分布概率最高的 Top20 特征词，64 个主题得到不重复的特征词表包括 657 个特征词。因此，每个主题表示为 657 维的特征词向量。基于算法 1 的流形学习步骤，将 64 个 657 维的主题向量映射在二维平面上。图 3 为 64 个主题的

ISOMAP 嵌入变量输出，每一个点代表一个主题，每一条边为 ISOMAP 构建的主题邻域图中主题间的连接边。

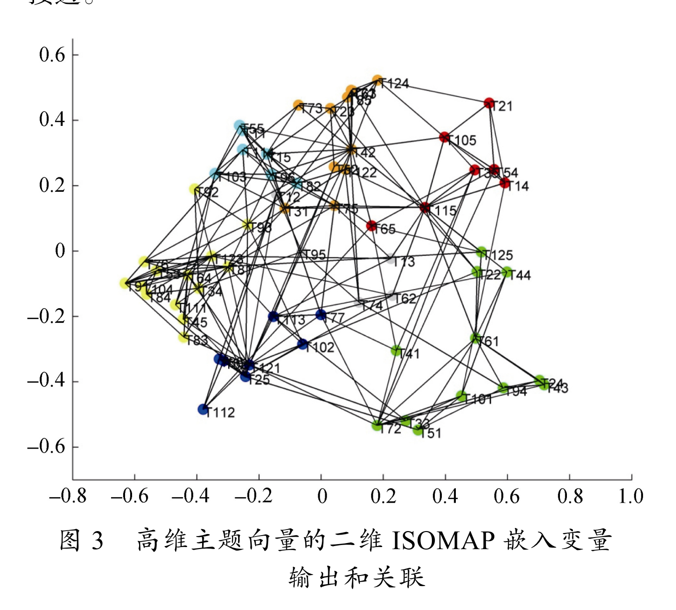


图 3 高维主题向量的二维 ISOMAP 嵌入变量输出和关联

通过分析发现主题在二维平面上的位置与该主题的特征词和内容相关。主题在二维平面上被聚类为 6 大类，分别为：黄色(军事)、青色(政治)、橘色(经济)、红色(科技)、蓝色(家庭/孩子)和绿色(生活)。例如，黄色节点标识的军事类主题，主要与南海领土问题、新疆恐怖主义、藏独、抗日战争胜利 70 周年大阅兵新闻报道相关；青色节点标识的政治类主题，主要与习主席与彭丽媛夫人出访、李克强总理访问、习主席与奥巴马总统会面、习主席访美等相关；橘色节点标识的经济类主题，主要与市场投资、中国股市泡沫、希腊债务相关；红色标识的社会科技类主题，主要与谷歌和小米等互联网公司、工业污染、波音飞机相关；蓝色节点标识的家庭/孩子类主题，主要与中国计划生育政策、二孩放开、孩子教育、张国立儿子吸毒等事件相关；绿色节点标识的生活类主题，与空气污染、柴静《穹顶之下》视频、优衣库试衣间视频等 2015 年引起媒体广泛讨论的民生事件相关。还有一些节点用灰色标注，这些节点较为分散，与上述 6 大节点簇距离较远。

可见，基于 ISOMAP 的非线性降维算法能够在低维嵌入平面正确表示主题之间的关联和相互作用关系；能够挖掘隐藏在高维向量间的规律、对相似的主题进行无监督学习聚类。即：基于 ISOMAP 的非线性降维算法对主题的聚类个数决定于主题向量之间的测地线距离，不需要根据先验知识事先确定，优于现有的依赖于算法初始值(如聚类个数和节点位置等)的聚类算法(如 K-means[22])。

4.3 主题演化结果分析

基于 3.4 节的方法，利用社会网络理论的度数(包括出度数和入度数)、介数中心度、密度和直径指标分析由 ISOMAP 算法得到的二维平面主题关联图。首先根据公式(10)为图 3 的每条边赋予权重，借鉴文献[4]的方法过滤权重值小于阈值的边(取阈值为 0.9)，并利用 Pajek 软件[23]描绘主题之间的演化关系如图 4 所示。其中，每一个节点代表一个主题，有线弧代表主题的演化方向。如经济类主题 T_{23} 和 T_{42} 之间有一条弧，表示从 T_{23} 演变到 T_{42}。

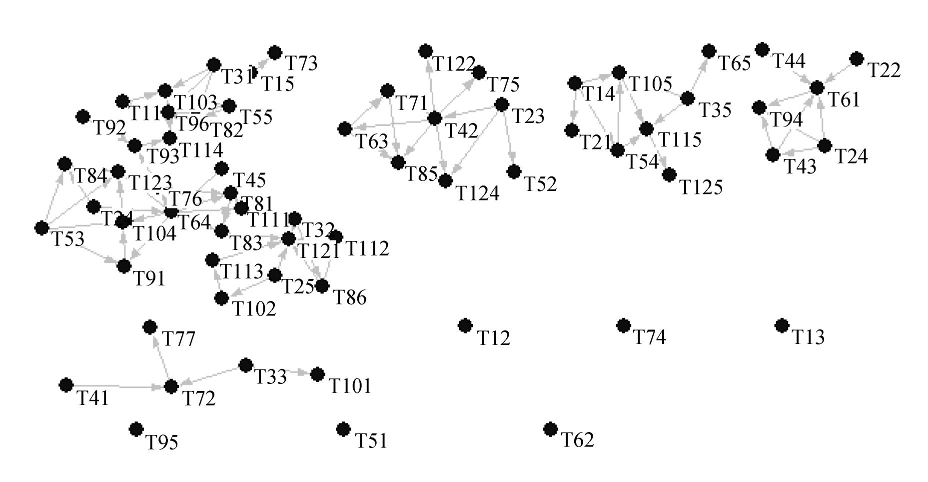


图 4 2015 年 CNN 对中国新闻报道的主题演化关系图

节点和有向弧构成了主题的演化路径，例如从图 4 的经济类主题中抽取一条路径为(T_{23}, T_{42}, T_{63}, T_{71})，其 Top3 主题特征词分别为 T_{23}(希腊、银行、债务)、T_{42}(投资、市场、增长)、T_{63}(市场、股票、投资)和 T_{71}(股票、经济危机、泡沫)。该路径的演变过程为: CNN 在 2015 年 2 月份对经济类主题的报道与中国和俄罗斯是否干预希腊债务有关，3 月份没有对经济类主题的报道，4 月、6 月和 7 月的经济类主题都与中国股票市场相关。由此可见，新闻主题的演变并不一定遵循相邻时间窗口的跨度，如 T_{23} 和 T_{42} 之间、T_{42} 和 T_{63} 之间。

注意到 5 月份有涉及经济类主题的报道(T_{52})，但并不在(T_{23}, T_{42}, T_{63}, T_{71})演化路径中，而是在另一个经济类主题的演化分支(T_{23}, T_{52})上，这是因为 T_{52} 除了涉及少量的中国股票市场泡沫的相关报道，主要涉及中俄经济、中国百万富翁增长等相关报道(参见表 3 列举的 5 月份 CNN 对中国经济所有相关报道的新闻标题)。若采用在相邻时间段中计算任意两个主题的特征向量相似度的方法，将会导致 T_{42} 和 T_{63} 之间关联出错，使得演化路径(T_{23}, T_{42}, T_{63}, T_{71})断裂。

根据图 4 的主题演化图计算每个主题的度数。表 4 为度数值最高和最低的 4 个主题。可以看出，度数值最高的 4 个主题是 T_{42}、T_{64}、T_{91}、T_{123}，主题的内容(见表 4 加黑标注的关键词)分别为股票增长/泡沫、南海军事安全、南海恐怖主义袭击、南海防御；这些主题在主题演化关系中局部影响力较高。而 T_{12}、T_{95}、T_{51} 和 T_{62} 为度数值最低(等于 0)的 4 个主题，此外还有 T_{74}、T_{13} 由于篇幅关系不一一列举。这些主题表现为孤立主题，大多为主题含义不明确(如 T_{12} 和 T_{95})或某个事件的突发报道(如 T_{51}，神州飞船发射)。

表 3　2015 年 5 月 CNN 与中国经济相关的所有新闻标题

新闻内容	时间	新闻标题
中俄经济	5 月 4 号	Russia and China have had enough of western banking.
中俄经济	5 月 11 号	China isn't Russia's answer to crisis with the West.
中国央行	5 月 19 号	China's central bank is just getting started.
中国首富	5 月 21 号	China's richest man lost $15 billion in one hour.
中国首富	5 月 22 号	China's richest man bet his company's shares would fall.
中国百万富翁	5 月 27 号	China has more than 1 million millionaires.
中国经济泡沫	5 月 31 号	The next big bubble: Bonds, startups, China?

表 4　度数值最高和最低的 4 个主题

主题	关键词(Top20)	入度	出度	度数
T_{42}	state, **investor**, government, **market**, **stock**, **growth**, power, global, charge, unite, **economic**, **trade**, **company**, washington, suspect, **bubble**, **bank**, **money**, president, department	9	1	10
T_{64}	government, attack, state, **sea**, **island**, **official**, hack, federal, information, **south**, unite, **security**, office, freedom, law, target, **cybersecurity**, **military**, personnel, international	6	3	9
T_{91}	**sea**, **official**, obama, issue, **island**, **cyber**, visit, **south**, **military**, xi, state, dispute, espionage, beijing, step, **attack**, tension, **security**, unite, **territorial**	3	3	6
T_{123}	state, **statement**, unite, **military**, pu, **island**, **defense**, complain, **sea**, job, freedom, **south**, economic, **dispute**, death, rule, criticize, fly, flight, post	0	6	6
T_{12}	musical, price, market, sun, sell, child, san, bao, baby, family, father, boy, bin, son, broadway, xiaomi, industry, police, production, man	0	0	0
T_{95}	panda, police, clip, bomb, glass, suspect, sprout, bridge, stock, sell, giant, trend, man, wednesday, kill, xinhua, money, thai, attack, Thursday	0	0	0
T_{51}	space, mission, shenzhou, astronaut, yang, kung, fu, opportunity, crewed, star, fei, wang, station, launch, man, war, center, return, zhang, nie	0	0	0
T_{62}	ship, yangtze, eastern, river, sink, star, cruise, state, capsize, rescue, water, passenger, june, authority, body, storm, board, tornado, monday, survivor	0	0	0

(注: 加黑标注的关键词能够清楚地反映主题的含义，因此重点标出。)

对于图 4 的主题演化图，计算每个主题的介数中心度。表 5 按照从高到低的顺序列举了介数中心度不为 0 的主题及其人工总结的主题内容。主题的介数中心度值越大，则在主题演化和关联关系中越活跃，媒介能力越强。可以看出，最活跃的主题为南海军事主题，其次为经济主题。而 7 月的优衣库主题(T_{72})、9 月

的纪念抗日战争胜利 70 周年的阅兵主题(T_{93})、9 月习主席出访美国华盛顿白宫主题(T_{96})、中国放开二孩政策主题(T_{32}、T_{86})和污染问题(T_{21})都是 2015 年度受到广泛关注、引起媒体热议的主题。由此可见，通过介数中心度指标能够找到主题演化关系中的活跃主题。

表 5　介数中心度值不为 0 的主题

主题	介数中心度	主题内容
T_{64}	0.00461	袭击 南海 安全
T_{91}	0.00307	南海 争议
T_{42}	0.00282	投资 市场 增长
T_{55}	0.00205	印度 交易 穆迪
T_{93}	0.00179	北京 阅兵 战争 习主席 军事
T_{96}	0.00166	习近平 美国 奥巴马 华盛顿 白宫
T_{72}	0.00154	优衣库 性 视频
T_{61}	0.00154	运动 伦敦 英国 足球 间谍 亚洲
T_{82}	0.00090	美国 习主席 奥巴马 货币贬值
T_{102}	0.00034	艺术 建筑师 比赛 网球 联赛
T_{21}	0.00026	工业 污染
T_{105}	0.00026	小米 手机 市场 科技 非洲
T_{113}	0.00026	北京 温度 冷 北韩 烟雾 零度以下
T_{84}	0.00026	弹道导弹 军事 检阅 军官 防御
T_{32}	0.00021	孩子 政策 人口数量
T_{86}	0.00021	孩子 家庭 父母 政策
T_{85}	0.00013	市场 股票 台湾 经济增长 金融风暴
T_{71}	0.00013	市场 股票 经济危机

对于图 4 的主题演化图，计算其网络拓扑图的密度为 0.02197266。密度值较小，网络的演变关系较为简单，与实际情况相符。

图 5 描绘了图 4 的主题演化图中最长的主题演化路径，长度为 5，为(T_{34}/T_{45}/T_{53}, T_{64}, T_{81}, T_{91}, T_{93}, T_{114})。这条路径描绘了 2015 年 CNN 媒体对我国军事和政治主题报道的演化过程，从 3 月、4 月份的徐才厚、周永康事件，到 5 月、6 月份的南海事件，到 8 月份中美讨论网络安全事件，再到 9 月份的纪念抗日战争胜利 70 周年阅兵事件，最后到 12 月份习主席和马英九在新加坡会面事件。

综合上述分析可得，实验结果与实际情况较为相符，可见基于流形学习的主题关系构建和演化分析方法能够在全局时间跨度建立主题的关联关系，挖掘主题关联关系间隐藏的规律并表征主题演化关系。该方法一方面克服高维主题特征向量之间的相似度计算带来的维数灾难问题，能够在低维平面输出主题的关联关系图，实现无监督的主题聚类和关联；另一方面避免了相邻时间窗口的主题关联失效而导致的全局主题演化链断裂，实验结果也表明新闻报道的演化并不遵循传统主题演化研究设定的相邻时间窗口跨度，而是具有不确定性和突发性；最后，基于社会网络相关指标能够较好地对新闻报道的主题演化结果进行分析和评价，找到主题演化过程中局部影响力较高的、较为活跃的主题。

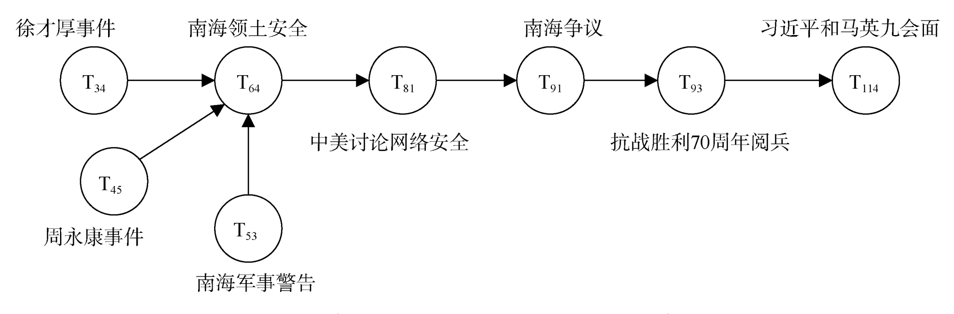


图 5　最长距离的主题演化路径示意图

5　结　语

本文提出一种基于流形学习的新闻主题关系构建和演化研究方法，利用流形学习在全局时间窗口对新闻领域的主题演化进行探索，通过对高维主题向量进行非线性降维并在低维空间重新定义话题间的距离以实现话题的关联，并借鉴社会网络理论的度数、介数中心度、密度和直径指标分析主题的演化结果。

以 2015 年美国 CNN 网站对中国相关的新闻报道为例对该方法的有效性进行验证，得出以下结论：非线性降维处理能够在低维嵌入平面正确表示主题之间的关联，并且能够挖掘隐藏在高维向量间的规律、实现对高维主题向量的约简和可视化；通过社会网络的度数和介数中心度指标能够找到话题演化关系中局部

影响力较大和较活跃的话题，通过密度和直径指标描绘整体的话题演化关系，并得到每一条主题演化路径。下一步工作是研究不同时间窗口长度对主题演化结果的影响以及基于可变时间窗口的主题演化分析。

参考文献:

[1] Samovar L A, Porter R E, McDaniel E R, et al. Communication Between Cultures [M]. Wadsworth, 2015.

[2] 楚克明, 李芳. 基于 LDA 模型的新闻主题的演化[J]. 计算机应用与软件, 2011, 28(4): 4-7, 26. (Chu Keming, Li Fang. LDA Model-based News Topic Evolution [J]. Computer Applications and Software, 2011, 28(4): 4-7, 26.)

[3] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[4] 楚克明. 基于LDA的新闻话题演化研究[D]. 上海: 上海交通大学, 2010. (Chu Keming. The Reaearch on Topic Evolution for News Based on LDA Model [D]. Shanghai: Shanghai Jiaotong University, 2010.)

[5] 胡艳丽, 白亮, 张维明. 一种话题演化建模与分析方法[J]. 自动化学报, 2012, 38(10): 1690-1697. (Hu Yanli, Bai Liang, Zhang Weiming. Modeling and Analyzing Topic Evolution [J]. Acta Automatic Sinica, 2012, 38(10): 1690-1697.)

[6] Seung H S, Lee D D. Cognition-The Manifold Ways of Perception [J]. Science, 2000, 290(5500): 2268-2269.

[7] Donoho D L. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality [C]. In: Proceedings of International Conference of Mathematicians, Paris, France. 2000: 6-11.

[8] Wang X, McCallum A. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends [C]. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006: 424-433.

[9] Blei D M, Lafferty J D. Dynamic Topic Models [C]. In: Proceedings of the 23rd International Conference on Machine Learning. 2006: 113-120.

[10] Hall D, Jurafsky D, Manning C D. Studying the History of Ideas Using Topic Models [C]. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2008: 363-371.

[11] 崔凯, 周斌, 贾焰, 等. 一种基于 LDA 的在线主题演化挖掘模型[J]. 计算机科学, 2010, 37(11): 156-159, 193. (Cui Kai, Zhou Bin, Jia Yan, et al. LDA-based Model for Online Topic Evolution Mining [J]. Computer Science, 2010, 37(11): 156-159, 193.)

[12] 李保利, 杨星. 基于 LDA 模型和话题过滤的研究主题演化分析[J]. 小型微型计算机系统, 2012, 33(12): 2738-2743. (Li Baoli, Yang Xing. Analyzing Research Topic Evolution with LDA and Topic Filtering [J]. Journal of Chinese Computer Systems, 2012, 33 (12): 2738-2743.)

[13] 秦晓慧, 乐小虬. 基于 LDA 主题关联过滤的领域主题演化研究[J]. 现代图书情报技术, 2015(3): 18-25. (Qin Xiaohui, Le Xiaoqiu. Topic Evolution Research on a Certain Field Based on LDA Topic Association Filter [J]. New Technology of Library and Information Service, 2015(3): 18-25.)

[14] Griffiths T L, Steyvers M. Finding Scientific Topics [J]. Proceedings of the National Academy Sciences of the United States of America, 2004, 101(1): 5228-5235.

[15] Cao J, Xia T, Li J. A Density-based Method for Adaptive LDA Model Selection [J]. Neurocomputing, 2009, 72(7-9): 1775-1781.

[16] Law M H C, Jain A K. Incremental Nonlinear Dimensionality Reduction by Manifold Learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(3): 377-391.

[17] Tenenbaum J B, De Silva V, Langford J C. A Global Geometric Framework for Nonlinear Dimensionality Reduction [J]. Science, 2000, 290(5500): 2319-2323.

[18] Roweis S T, Saul L K. Nonlinear Dimensionality Reduction by Locally Linear Embedding [J]. Science, 2000, 290(5500): 2323-2326.

[19] Maning C D, Schütze H, Raghavan P. 信息检索导论[M]. 王斌译. 北京: 人民邮电出版社, 2011. (Manning C D, Schütze H, Raghavan P. Introduction to Information Retrieval [M]. Translated by Wang Bin. Beijing: Post &Telecom Press, 2011.)

[20] Costa L, Da F, Rodrigues F A, et al. Characterization of Complex Networks: A Survey of Measurements [J]. Advances in Physics, 2007, 56(1): 167-242.

[21] GooSeeker [EB/OL]. http://www.gooseeker.com.

[22] Hartigan J A, Wong M A. Algorithm AS: A K-means Clustering Algorithm [J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1979, 28(1): 100-108.

[23] Pajek: Analysis and Visualization of Large Networks [EB/OL]. http://mrvar.fdv.uni-lj.si/pajek/.

作者贡献声明:

徐月梅: 提出研究思路, 设计研究方案, 撰写论文;
李杨: 设计研究方案, 全文修改定稿;

梁野: 采集、清洗和分析数据;
蔡连侨: 提出部分修改意见。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: xuyuemei@bfsu.edu.cn。

[1] 徐月梅, 李杨, 梁野, 蔡连侨. ISOMAP.mat. ISOMAP 主题降维算法 Matlab 程序.

[2] 徐月梅, 李杨, 梁野, 蔡连侨. Cnn_China_2015.xlsx. 2015 年美国有线电视新闻网与中国相关的新闻.

[3] 徐月梅, 李杨, 梁野, 蔡连侨. data.txt. 预处理后的数据集.

[4] 徐月梅, 李杨, 梁野, 蔡连侨. LDA.mat. LDA 主题抽取算法的 Matlab 程序.

[5] 徐月梅, 李杨, 梁野, 蔡连侨. LDAresult.xlsx. LDA 抽取得到的主题和主题关键词.

[6] 徐月梅, 李杨, 梁野, 蔡连侨. modelresult.xlsx. 主题降维后各主题在低维平面的坐标和关联边.

[7] 徐月梅, 李杨, 梁野, 蔡连侨. Similar.mat. 低维平面关联边的权重计算算法.

[8] 徐月梅, 李杨, 梁野, 蔡连侨. Pajekinput.net. Pajek 软件生成主题演化图的数据.

收稿日期: 2016-05-13
收修改稿日期: 2016-08-23

Analyzing Evolution of News Topics with Manifold Learning

Xu Yuemei[1] Li Yang[2, 3] Liang Ye[1] Cai Lianqiao[1]
[1](Department of Computer Science, Beijing Foreign Studies University, Beijing 100089, China)
(Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)
[3](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] This study aims to examine the creation and development of online news topics, and then to gauge the public opinion. [**Methods**] First, we introduced the manifold learning technology to analyze the news topics. Second, we explored the relations among the high dimensional topics from each time window, which were identified by the LDA model. Third, we clustered these topics and visualized the relations among them in the low-dimensional space. Finally, we analyzed the topic evolution with the help of social network theorem. [**Results**] The proposed method could effectively identify the topic evolution trends of news reports on China from CNN in 2015. [**Limitations**] We did not fully explore the impacts of time windows. [**Conclusions**] This study provides a new method to visualize the evolution of news report topics over a period of time, which avoids inaccurate description due to the changing of adjacent time windows.

Keywords: Latent Dirichlet Allocation Manifold learning Topic relevance Topic evolution